

Weighted envelope estimation to handle variability in model selection

Daniel J. Eck and R. Dennis Cook

January 5, 2017

Abstract

Envelope methodology can provide substantial efficiency gains in multivariate statistical problems, but in some applications the estimation of the envelope dimension can induce selection volatility that will mitigate those gains. Current envelope methodology does not account for the added variance that can result from this selection volatility. In this article, we circumvent dimension selection volatility through the development of a weighted envelope estimator. Theoretical justification is given for our weighted envelope estimator and validity of the residual bootstrap approximation for the multivariate regression model is established. A simulation study and an analysis on a real data set illustrate the utility of our weighted envelope estimator.

Keywords: Dimension Reduction; Envelope Models; Model Selection; Residual Bootstrap; Variance Reduction.

1 Introduction

Envelope methodology was developed originally in the context of the multivariate linear regression model (Cook, et al., 2010),

$$Y = \alpha + \beta X + \varepsilon, \quad (1)$$

where $\alpha \in \mathbb{R}^r$, the random response vector is $Y \in \mathbb{R}^r$, the fixed predictor vector $X \in \mathbb{R}^p$ is centered to have mean zero, and the error vector $\varepsilon \sim N(0, \Sigma)$. It was shown by Cook, et al. (2010) that the envelope estimator of the unknown coefficient matrix $\beta \in \mathbb{R}^{r \times p}$ in (1) has the potential to yield massive efficiency gains relative to the standard estimator of β . These efficiency gains can arise when the dimension u of the envelope, defined in the next section, is less than r . In most practical applications, u is unknown and has to be estimated. This estimation can be problematic since the estimated variance of the envelope estimator is typically calculated conditional on the estimated dimension u . Variation associated with model selection is therefore not considered in the current envelope paradigm.

In this article, we propose a weighted envelope estimator of β that smooths out model selection volatility. The weighting is across all possible envelope models under (1). The weights corresponding to each envelope estimator are functions of the Bayesian Information Criterion (BIC) value corresponding to that particular envelope model. Weighting in this manner is similar to the model averaging techniques discussed by Buckland, et al.

(1997) and Burnham and Anderson (2004) who provided a philosophical justification for the use of such weighted estimators without giving any theoretical properties. Hjort and Claeskens (2003) and Liang, et al. (2011) built on the framework of Buckland, et al. (1997) and Burnham and Anderson (2004) by deriving the asymptotic properties for weighted estimators of generalized linear regression parameters with weighting conducted across submodels under consideration.

2 The Envelope Model

The original motivation for envelope methodology comes from the observation that, in the multivariate regression model (1), some linear combinations of Y may have a distribution that does not depend on X , while other linear combinations of Y do depend on X . The envelope model separates out these immaterial and material parts of Y , and thereby allows for efficiency gains (Cook, et al., 2010; Su and Cook, 2011).

More carefully, suppose that we can find a subspace $\mathcal{S} \subseteq \mathbb{R}^r$ so that

$$\mathcal{Q}_{\mathcal{S}}Y \perp \mathcal{P}_{\mathcal{S}}Y|X, \quad \text{and} \quad \mathcal{Q}_{\mathcal{S}}Y|X = x_1 \sim \mathcal{Q}_{\mathcal{S}}Y|X = x_2, \quad \text{for all } x_1, x_2, \quad (2)$$

where \sim means identically distributed, $\mathcal{P}_{(\cdot)}$ projects onto the subspace indicated by its argument and $\mathcal{Q} = I_r - \mathcal{P}$. For any \mathcal{S} with the properties (2), $\mathcal{P}_{\mathcal{S}}Y$ carries all of the material information and perhaps some of the immaterial information, while $\mathcal{Q}_{\mathcal{S}}$ contains just immaterial information. Let $\mathcal{B} = \text{span}(\beta)$. Then (2) holds if and only if $\mathcal{B} \subseteq \mathcal{S}$ and $\Sigma = \Sigma_{\mathcal{S}} + \Sigma_{\mathcal{S}^\perp}$, where $\Sigma_{\mathcal{S}} = \text{var}(\mathcal{P}_{\mathcal{S}}Y)$ and $\Sigma_{\mathcal{S}^\perp} = \text{var}(\mathcal{Q}_{\mathcal{S}}Y)$. The envelope is defined as the intersection of all subspaces \mathcal{S} that satisfy (2) and is denoted by $\mathcal{E}_{\Sigma}(\mathcal{B})$ with dimension $u = \dim\{\mathcal{E}_{\Sigma}(\mathcal{B})\}$.

The envelope model can be represented in terms of coordinates by parameterizing model (1) to incorporate conditions (2). Define $\Gamma \in \mathbb{R}^{r \times u}$ to be a semi-orthogonal basis matrix for $\mathcal{E}_{\Sigma}(\mathcal{B})$ and let $\Gamma_o \in \mathbb{R}^{r \times (r-u)}$ be a completion of Γ so that $(\Gamma, \Gamma_o) \in \mathbb{R}^{r \times r}$ is an orthogonal matrix. Then the envelope model with respect to model (1) is parameterized as

$$Y = \alpha + \Gamma\eta X + \varepsilon, \quad \varepsilon \sim N(0, \Sigma), \quad (3)$$

where $\Sigma = \Gamma\Omega\Gamma^T + \Gamma_o\Omega_o\Gamma_o^T$, $\Omega \in \mathbb{R}^{u \times u}$ and $\Omega_o \in \mathbb{R}^{(r-u) \times (r-u)}$ are positive definite, and $\eta \in \mathbb{R}^{u \times p}$ is β in the coordinates of Γ . We see from (3), that $\mathcal{E}_{\Sigma}(\mathcal{B})$ links the mean and covariance structures of the regression problem and it is this link that provides the efficiency gains. The gains can be massive when the immaterial information is large relative to the material information; for instance, when $\|\Omega\| \ll \|\Omega_o\|$, where $\|\cdot\|$ is a matrix norm (Cook, et al., 2010). An illuminating schematic showing how an envelope increases efficiency was given by Su and Cook (2011).

Candidate envelope estimators of β at dimension j and sample size n , denoted $\hat{\beta}_j$, are found via maximum likelihood estimation of model (3) with $\hat{\beta}_j = \widehat{\Gamma}\hat{\eta}$. The envelope estimator of β is found by comparing all candidate envelope estimators using a model selection criterion such as BIC, or likelihood ratio tests or perhaps cross validation. The estimated dimension, \hat{u} , obtained from any one of these selection criteria is a variable quantity dependent on the observed data. Traditional envelope methodology does not address this extra variability. In the next three sections, we develop new envelope methodology that takes this extra variability into account.

3 BIC Weighted Estimators

We develop a solution to the problem of potential volatility in envelope model selection by building on the ideas in Buckland, et al. (1997) and Burnham and Anderson (2004), who suggested combining estimators over different models by weighting. Bootstrapping was then suggested for stochastic weighting schemes, but no theoretical properties were given by the authors.

We consider weighted estimators of the form

$$\hat{\beta}_w = \sum_{j=1}^r w_j \hat{\beta}_j, \quad (4)$$

where $\sum_{j=1}^r w_j = 1$ and $w_j \geq 0$, for $j = 1, \dots, r$. The weights w_j depend on the BIC values for all of the candidate envelope models under consideration. Let the BIC value for the envelope model with dimension j be denoted by $b_j = -2l(\hat{\beta}_j) + k(j) \log(n)$, where $l(\hat{\beta}_j)$ is the log likelihood evaluated at the envelope estimator $\hat{\beta}_j$ and $k(j)$ is the number of parameters of the envelope model of dimension j . The weights for envelope model j are constructed as

$$w_j = \frac{\exp(-b_j)}{\sum_{k=1}^r \exp(-b_k)}. \quad (5)$$

It follows from the Supplement that $\hat{\beta}_w$ is a \sqrt{n} -consistent estimator of β , but assessing the variance of $\hat{\beta}_w$ is not so straightforward. In the next section we show that the residual bootstrap provides a consistent estimator of $\text{var}(\hat{\beta}_w)$.

Similar weights corresponding to Akaike's Information Criterion (AIC) do not have the nice asymptotic properties that weights corresponding to BIC enjoy. In particular, analogous AIC weight at $j = u$ is not guaranteed to converge to 1 asymptotically. Additionally, the weights in (5) differ slightly from those mentioned in Burnham and Anderson (2004) which were also advocated by Kass and Raftery (1995) and Tsague (2014). These weights are of the form

$$\tilde{w}_j = \frac{\exp(-b_j/2)}{\sum_{k=1}^r \exp(-b_k/2)} \quad (6)$$

and they correspond to an approximation of the posterior probability for model j given the observed data under the prior which places equal weight for all candidate models. Weights of the form (6) do not have the same asymptotic properties as the weights given by (5). A more thorough discussion of this is given after Theorem 1.

4 Bootstrap for $\hat{\beta}_w$

The residual bootstrap used to estimate the variability for the envelope estimator at the true dimension u uses the starred responses,

$$Y^* = \mathbb{X} \hat{\beta}_u^T + \varepsilon^*, \quad (7)$$

to obtain $\hat{\beta}_u^*$, where $\mathbb{X} \in \mathbb{R}^{n \times p}$ is the fixed design matrix with rows X_i and the rows of ε^* are the realizations of n resamples of the residuals from the original model fit with replacement. The envelope estimator $\hat{\beta}_u$ is \sqrt{n} -consistent and asymptotically normal (Cook, et al., 2010; Cook and Zhang, 2015). The techniques used to verify the consistency and asymptotic normality of $\hat{\beta}_u$ require the asymptotics of extremum estimation as in Amemiya

(1985, Theorems 4.1.1-4.1.3). The setup in Andrews (2002, Section 2 pgs. 122-124 and Theorem 2) confirms that the residual bootstrap, with responses (7), provides a \sqrt{n} -consistent estimator of the asymptotic variability of $\hat{\beta}_u$. The problem with this approach, as it currently stands, is that u is unknown. The current implementation of the residual bootstrap implicitly assumes that $\hat{u} = u$ where \hat{u} is obtained via some selection criterion. Therefore, variability introduced by model selection uncertainty is ignored. This issue is resolved by using $\hat{\beta}_w$ in place of $\hat{\beta}_u$ in (7). The next theorem formalizes our asymptotic justification for the use of the weighted envelope estimator $\hat{\beta}_w$ in practical problems. Its proof is given in the Supplement.

Theorem 1. *Assume the regression model (1) and suppose that an envelope subspace of dimension $u = 1, \dots, r$ exists. Assume that $\frac{1}{n}\mathbb{X}^T\mathbb{X} \rightarrow \Sigma_X > 0$. Let $\hat{\beta}_w$ be the weighted envelope estimator of β defined in (4) and let $\hat{\beta}_w^*$ be the weighted envelope estimator of β obtained from resampled data. Then, as n tends to ∞ ,*

$$\begin{aligned} \sqrt{n} \{ \text{vec}(\hat{\beta}_w^*) - \text{vec}(\hat{\beta}_w) \} &= \sqrt{n} \{ \text{vec}(\hat{\beta}_u^*) - \text{vec}(\hat{\beta}_u) \} \\ &+ O_p \{ n^{(1/2-p)} \} + 2(u-1)O_p(1)\sqrt{n}e^{-n|O_p(1)|}. \end{aligned} \quad (8)$$

Theorem 1 shows the utility of the weighted envelope estimator $\hat{\beta}_w$. In (8), we see that asymptotic distribution of the residual bootstrap with respect to $\hat{\beta}_w$ is the same as the asymptotic distribution of the residual bootstrap at $\hat{\beta}_u$, the envelope estimator at the true dimension. The difference between the two bootstrap procedures is that the bootstrap given in Theorem 1 does not require the conditioning on \hat{u} as a prerequisite for its implementation. We instead bootstrap with respect to a tangible estimator that does not ignore key elements of variability that are apparent in practical problems.

The orders in (8) result from model selection variability that arises from four sources. The $O_p \{ n^{(1/2-p)} \}$ term corresponds to the rate at which $\sqrt{n}w_j$ and $\sqrt{n}w_j^*$ vanish for $j = u+1, \dots, r$. This rate is a cost of over estimation of the envelope space. It decreases quite fast, particularly when p is not small, because models with $j > u$ are true and thus have no systematic bias due to choosing the wrong dimension.

The $2(u-1)\sqrt{n}e^{-n|O_p(1)|}$ term corresponds to the rate at which $\sqrt{n}w_j$ and $\sqrt{n}w_j^*$ vanish for $j = 1, \dots, u-1$. This rate arises from under estimating the envelope space and it is affected by systematic bias arising from choosing the wrong dimension. To gain intuition about this rate, let $B_j = (G_o^T \Sigma G_o)^{-1/2} G_o^T \beta \Sigma_X^{1/2}$, where $G_o \in \mathbb{R}^{r \times (r-j)}$ is the population basis matrix for the complement of the envelope space of dimension j . This quantity is a standardized version of $G_o^T \beta$ that reflects bias, since $G_o^T \beta \neq 0$ when $j < u$, but $G_o^T \beta = 0$ when $j \geq u$. Let $\hat{B}_{j,n}$ denote the \sqrt{n} -consistent estimator of B_j obtained by plugging in the sample version of Σ_X and the estimators of G_o , Σ and β that arise by maximizing the likelihood with dimension $j < u$. Then the $-n | O_p(1) |$ term appearing in the exponent of $2(u-1)\sqrt{n}e^{-n|O_p(1)|}$ is the rate at which $-n \log(|I_p + \hat{B}_{j,n}^T \hat{B}_{j,n}|)$ approaches $-\infty$. Additionally, this term is 0 when $u = 1$. That arises because we consider only regressions in which $\beta \neq 0$ and thus $u \geq 1$. When $u = 1$ underestimation is not possible in our context and thus $2(u-1)\sqrt{n}e^{-n|O_p(1)|}$ vanishes.

We now revisit the origins of construction of the weights used in Theorem 1. In Section 3, we mentioned that our construction is similar to, but not the same as, those mentioned in Burnham and Anderson (2004). In the case when $p = 1$, the term $\sqrt{n}\tilde{w}_{j=u+1}$ defined by (6) does not vanish as $n \rightarrow \infty$. We therefore would not have the same asymptotic result given by (8) in Theorem 1. Instead, there would be non-zero weight placed on the

envelope model with dimension $j = u + 1$ asymptotically. This weighting scheme would therefore lead to higher estimated variability than is necessary in practice. However, this issue is no longer problematic when $p > 1$. When $p > 1$, the weights (6) can be used and changes to (8) would result. The $O_p \{n^{1/2-p}\}$ term in (8) would become $O_p \{n^{(1-p)/2}\}$ when the weights (6) are used in place of the weights (5). When p is large, one may proceed with weighting according to (6) at relatively little cost to efficiency.

5 Examples

We now provide examples which show that our weighted envelope estimator performs better than the standard estimator and favorably with other envelope estimators at reasonable values of u . The first two are simulated examples in which we know β , Σ , u , and $\mathcal{P}_{\mathcal{E}_\Sigma(B)}$. Their role is only to illustrate the theory developed in the previous sections. The third example is a real data example in which we do not know any quantities of interest.

5.1 Simulated examples

Example 1: For this example, we create a setting in which $Y \in \mathbb{R}^3$ is generated according to the model

$$Y_i = \beta_i X_i + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{ind}}{\sim} N(0, \Sigma), \quad (9)$$

$i = 1, \dots, n$, where $X_i \in \mathbb{R}^2$ is a continuous predictor with entries generated independently from a normal distribution with mean 4 and variance 1. The covariance matrix Σ was generated using three orthonormal vectors and has eigenvalues of 50, 10, and 0.01. The matrix $\beta \in \mathbb{R}^{3 \times 2}$ is an element in the space spanned by the second and third eigenvectors of Σ . We know that the dimension of $\mathcal{E}_\Sigma(B)$ is $u = 2$. Three datasets were simulated using model (9) at different sample sizes, as given in Table 1. The multivariate residual bootstrap was then used to compare the efficiencies of our weighted envelope estimator $\hat{\beta}_w$ to the oracle envelope estimator $\hat{\beta}_{u=2}$. The ratios of bootstrapped estimated standard errors between both envelope estimators to those of the maximum likelihood estimator (MLE) from the full model, $\text{se}(\hat{\beta}_r)/\text{se}^*(\hat{\beta}_w)$, are seen in Table 1. Ratios greater than 1 indicate that the envelope estimator is more efficient than the standard estimator. There are two conclusions that are apparent from Table 1. We see that envelope estimation is more efficient than the estimation using the full model and we see that the efficiency of the weighted envelope estimator approaches that of the oracle estimator, $\hat{\beta}_{u=2}$, as n increases.

Example 2: For this example, we illustrate the effect that p has on the performance of the weighted envelope estimator. We generated data according to model (9) with $Y \in \mathbb{R}^5$. In this example $u = 1$ and Σ is compound symmetric with diagonal entries set to 1 and off-diagonal entries set to 0.5, $\beta = 1_r c_p^T$, where 1_r is the $r \times 1$ vector of ones, c_p is a $p \times 1$ vector where every entry is 10. We generate the predictors according to $X \sim N(0, I_p)$, where I_p is the p -dimensional identity matrix. We set $n = 250$.

The results of our simulation study are seen in Table 2. For each value of p that is considered, we display the number of estimated dimensions \hat{u} as determined by BIC. From Table 2, we see that the distribution of \hat{u} approaches a point mass at the truth as p increases. This implies that the bias terms in Theorem 1 vanish as p increases just as (8) states.

$n = 250$		$n = 500$		$n = 2000$	
$\hat{\beta}_w$	$\hat{\beta}_{u=2}$	$\hat{\beta}_w$	$\hat{\beta}_{u=2}$	$\hat{\beta}_w$	$\hat{\beta}_{u=2}$
1.88	2.40	2.34	2.98	2.71	2.81
1.39	1.79	1.65	1.78	1.79	1.81
2.67	3.60	2.57	3.52	3.51	3.71
2.33	2.66	2.18	2.99	2.67	2.79
1.87	1.86	1.67	1.81	1.73	1.77
3.39	3.75	2.52	3.70	3.36	3.74

Table 1: Ratios of estimated standard errors obtained from the multivariate residual bootstrap for a different number of sample sizes n .

	$n(\hat{u} = 1)$	$n(\hat{u} = 2)$	$n(\hat{u} = 3)$
$p = 2$	128	111	11
$p = 5$	214	34	2
$p = 10$	249	1	0
$p = 25$	250	0	0

Table 2: Simulation results for Example 2.

5.2 Cattle data

The data for this illustration resulted from an experiment to compare two treatments for the control of an intestinal parasite in cattle: thirty animals were randomly assigned to each of the two treatments and their weights (in kilograms) were recorded at weeks 2, 4,..., 18 and 19 after treatment (Kenward, 1987). Because of the nature of a cows digestive system, the treatments were not expected to have an immediate measurable affect on weight. The objectives of the study were to find if the treatments had differential effects on weight and, if so, about when were they first manifested. We begin by consider the multivariate linear model (1), where $Y_i \in \mathbb{R}^{10}$ is the vector of cattle weights from week 2 to week 19, and the binary predictor X_i is either 0 or 1 indicating the two treatments. Then $\alpha = E(Y|X = 0)$ is the mean profile for one treatment and $\beta = E(Y|X = 1) - E(Y|X = 0)$ is the mean profile difference between treatments.

Turning to a fit of the envelope model (3), likelihood ratio testing selects $\hat{u} = 1$ and BIC selects $\hat{u} = 3$ as the dimension of the envelope model. Further complicating matters, when BIC is used to determine u at every iteration of the multivariate residual bootstrap, we see high variability in model selection as seen in Table 3. From Table 3, it appears that the true dimension of the envelope subspace is anywhere from 1 to 5 with the highest likelihood that it is between 2 and 4. Model selection volatility of this variety is precisely the reason why the weighted envelope estimator is advocated; it would not be safe to perform a bootstrap procedure that makes a uniform selection of a particular dimension at every iteration. Such a procedure ignores the model selection variability seen in Table 3.

From Table 4, we see the ratios of bootstrapped estimated standard errors between both envelope estimators to

those of the MLE from the full model, $\text{se}(\hat{\beta}_r)/\text{se}^*(\hat{\beta}_w)$. Ratios greater than 1 indicate that the envelope estimator is more efficient than the standard estimator. We see that $\hat{\beta}_w$ is comparable to $\hat{\beta}_{u=3}$. Similar conclusions are drawn from the other elements of estimates of β . The findings displayed in Table 4 show that the weighted envelope estimator can provide useful efficiency gains while protecting against underestimation of u that may not be properly account for by the standard envelope estimator .

\hat{u}	1	2	3	4	5
$n(\hat{u})$	10	10	24	12	4

Table 3: Counts of the selected envelope dimension at every iteration of a multivariate residual bootstrap for 60 resamples.

d	B	$\hat{\beta}_w$	$\hat{\beta}_{u=1}$	$\hat{\beta}_{u=2}$	$\hat{\beta}_{u=3}$	$\hat{\beta}_{u=4}$	$\hat{\beta}_{u=5}$
5	60	1.93	4.65	3.89	1.85	1.54	1.27
	100	1.38	3.97	1.49	1.14	1.14	1.07
	200	1.62	4.26	3.14	1.69	1.32	1.19
	500	1.61	4.58	2.43	1.59	1.29	1.15
	1000	1.56	4.10	2.48	1.55	1.29	1.15
	2000	1.57	4.43	2.30	1.53	1.28	1.16
6	60	1.75	2.30	2.35	1.79	1.38	1.24
	100	1.25	2.15	1.26	1.05	1.05	1.00
	200	1.50	2.27	2.47	1.55	1.20	1.11
	500	1.50	2.22	2.05	1.55	1.24	1.10
	1000	1.52	2.24	1.99	1.48	1.26	1.14
	2000	1.53	2.32	1.91	1.46	1.26	1.16

Table 4: Ratios of estimated standard errors obtained from the multivariate residual bootstrap at a different number of resamples B for the fifth and sixth elements (indicated by the d column) of estimates of β .

6 Discussion

Efron (2014) proposed an estimator motivated by bagging (Breimen, 1996) that aims to reduce variability and smooth out discontinuities resulting from model selection volatility. Variability of the model averaged estimator of Efron (2014) is assessed via a double bootstrap. These techniques have been applied to envelope methodology in Eck, et al. (2016) and useful variance reduction was found empirically. The problem of interest in Eck, et al. (2016) falls outside the scope of the multivariate linear regression model, and general envelope methodology (Cook and Zhang, 2015) was required to obtain efficiency gains. In the context of the multivariate linear regression model, we show that only a single level of bootstrapping is necessary to assess the variability of our weighted

envelope estimator and that bootstrapping in this way guarantees a consistent estimator of the variability of the weighted envelope estimator.

7 Supplementary material

Supplementary material available at Biometrika online includes the proof of Theorem 1.

References

- Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge, MA.
- Andrews, D. W. K. (2002). Higher-Order Improvements of a Computationally Attractive k -Step Bootstrap for Extremum Estimators. *Econometrica*, **70**, 1, 119–162.
- Breiman, L. (1996). Bagging Predictors. *Machine Learning*, **24**, 123–140.
- Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model Selection: An Integral Part of Inference. *Biometrics*, **53**, 603–618.
- Burnham, K. P., Anderson, D. R. (2004). Multimodel Inference. *Sociological and Methods Research*, **33**, 261–304
- Cook, R. D., Li, B., Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, **20**, 927–1010.
- Cook, R. D., Forzani, L., and Su, Z. (2016). A note on fast envelope estimation. *J. Mult. Anal.*, **150**, 42–54.
- Cook, R. D., Zhang, X. (2015). Foundations for Envelope Models and Methods. *J. Am. Statist. Assoc.*, **110:510**, 599–611.
- Eck, D. J., Geyer, C. J., and Cook, R. D. (2016). An Application of Envelope and Aster Models. *Submitted*.
- Efron, B. (2014). Estimation and Accuracy After Model Selection. *J. Am. Statist. Assoc.*, **109:507**, 991–1007.
- Hjort, N. L. and Claeskens, G. (2003). Frequentist Model Average Estimators. *J. Am. Statist. Assoc.*, **98:464**, 879–899.
- Kass, R. K. and Raftery, A. E. (1995). Bayes Factors. *J. Am. Statist. Assoc.*, **90:430**, 775–795.
- Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. *J. R. Statist. Soc. C*, **36**, 296–308.
- Liang, H., Zou, G., Wan, A. T. K., and Zhang, X. (2011). Optimal Weight Choice for Frequentist Model Average Estimators. *J. Am. Statist. Assoc.*, **106:495**, 1053–1066.
- Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika*, **98**, 133–146.

Tsague, G. N. (2014). On Optimal Weighting Scheme in Model Averaging. *American Journal of Applied Mathematics and Statistics*, **2**, No. 3, 150–156.

‘Supplementary material for Weighted envelope estimation to handle variability in model selection’

This Supplementary Materials section contains the proof of Theorem 1 in Eck and Cook (2017).

Proof. We go through the steps showing that (8) in Eck and Cook (2017) holds. Recall that $u = \dim(\mathcal{E})$. Define $l(\hat{\beta}_j)$ to be the log likelihood of the envelope model evaluated at the envelope estimator $\hat{\beta}_j$, fitting with $\dim(\mathcal{E}) = j$, and define $k(j)$ to be the number of parameters of the envelope model of dimension j . From the construction of b_j and the above calculations we see that

$$e^{b_u - b_j} = e^{-2\{l(\hat{\beta}_u) - l(\hat{\beta}_j)\}} n^{-\{k(j) - k(u)\}}.$$

Let b_j^* be the BIC value of the envelope model of dimension j fit to the starred data and define

$$w_j^* = \frac{e^{-b_j^*}}{\sum_{k=1}^r e^{-b_k^*}}.$$

Let $\|\cdot\|$ be the Euclidean norm. We show that $\sqrt{n}\{w_j^* \text{vec}(\hat{\beta}_j^*) - w_j \text{vec}(\hat{\beta}_j)\} \rightarrow 0$ for $j \neq u$ by showing that

$$\sqrt{n}\|w_j^* \text{vec}(\hat{\beta}_j^*) - w_j \text{vec}(\hat{\beta}_j)\| \leq \sqrt{n}\|w_j^* \text{vec}(\hat{\beta}_j^*)\| + \sqrt{n}\|w_j \text{vec}(\hat{\beta}_j)\| \rightarrow 0$$

as $n \rightarrow \infty$ for all $j \neq u$. Now,

$$\begin{aligned} \sqrt{n}w_j\|\text{vec}(\hat{\beta}_j)\| &\leq \sqrt{n} |O_p(1)| e^{b_u - b_j} \\ &= |O_p(1)| n^{\{k(u) - k(j) + 1/2\}} e^{-2\{l(\hat{\beta}_u) - l(\hat{\beta}_j)\}} \\ &= |O_p(1)| n^{\{k(u) - k(j) + 1/2\}} e^{2\{l(\hat{\beta}_r) - l(\hat{\beta}_j)\} - 2\{l(\hat{\beta}_r) - l(\hat{\beta}_u)\}}. \end{aligned} \tag{10}$$

The first inequality in (10) follows from the fact that $\|\text{vec}(\hat{\beta}_j)\| \leq \|\text{vec}(\hat{\beta}_r)\|$ and $\|\text{vec}(\hat{\beta}_r)\| = O_p(1)$. We first consider the case where $j = u + 1, \dots, r$. In this setting, models with envelope dimensions u and j are both true and nested within the full model with envelope dimension r . Consequently, $-2\{l(\hat{\beta}_u) - l(\hat{\beta}_r)\}$ and $-2\{l(\hat{\beta}_j) - l(\hat{\beta}_r)\}$ are asymptotically distributed as $\chi_{p(r-u)}^2$ and $\chi_{p(r-j)}^2$ by Wilks' Theorem. Therefore $e^{-2\{l(\hat{\beta}_u) - l(\hat{\beta}_j)\}} = O_p(1)$ since it is the exponentiation of the difference between two χ^2 random variables. We see that

$$\sqrt{n}w_j\|\text{vec}(\hat{\beta}_j)\| \leq |O_p(1)| n^{\{k(u) - k(j) + 1/2\}} = O_p[n^{\{k(u) - k(j) + 1/2\}}].$$

Since $j > u$, we have that $k(u) - k(j) = p(u - j) \leq -p$. Thus,

$$\sqrt{n}w_j\|\text{vec}(\hat{\beta}_j)\| \leq O_p\{n^{(1/2-p)}\}$$

for $j = u + 1, \dots, r$. Following the same steps as (10), applied to the starred data, yields

$$\sqrt{n}w_j^*\|\text{vec}(\hat{\beta}_j^*)\| \leq |O_p(1)| n^{\{k(u) - k(j) + 1/2\}} e^{-2\{l^*(\hat{\beta}_u^*) - l^*(\hat{\beta}_r^*)\} + 2\{l^*(\hat{\beta}_j^*) - l^*(\hat{\beta}_r^*)\}} \tag{11}$$

where $l^*(\cdot)$ is the log likelihood function corresponding to the starred data. Both $-2\{l^*(\hat{\beta}_u^*) - l^*(\hat{\beta}_r^*)\}$ and $2\{l^*(\hat{\beta}_j^*) - l^*(\hat{\beta}_r^*)\}$ in (11) are $O_p(1)$. Thus,

$$\sqrt{n}w_j\|\text{vec}(\hat{\beta}_j^*)\| \leq |O_p(1)| n^{\{k(u) - k(j) + 1/2\}} = O_p[n^{\{k(u) - k(j) + 1/2\}}],$$

and, $\sqrt{n}w_j \|\text{vec}(\hat{\beta}_j^*)\| \leq O_p \{n^{(1/2-p)}\}$ for all $j = u+1, \dots, r$. This establishes that

$$\sqrt{n} \|w_j^* \text{vec}(\hat{\beta}_j^*) - w_j \text{vec}(\hat{\beta}_j)\| \leq O_p [n^{\{1/2-p\}}],$$

for $j = u+1, \dots, r$.

Turning to the case when $j = 1, \dots, u-1$, consider the exponent $e^{-\lambda_j}$, with $\lambda_j = 2 \{l(\hat{\beta}_r) - l(\hat{\beta}_j)\}$. This is a log likelihood ratio although, unlike the case when $j = u+1, \dots, r$, it does not follow a χ^2 distribution asymptotically. Let \widehat{G} and \widehat{G}_o be the estimated bases for the envelope space and its orthogonal complement fitting with dimension $j = 1, \dots, u-1$, so $\widehat{G} \in \mathbb{R}^{r \times j}$ and $\widehat{G}_o \in \mathbb{R}^{r \times (r-j)}$. We write

$$\begin{aligned} \lambda_j &= 2 \{l(\hat{\beta}_r) - l(\hat{\beta}_j)\} \\ &= n \log |\widehat{G}^T \widehat{\Sigma}_{\text{res}} \widehat{G}| + n \log |\widehat{G}_o^T \widehat{\Sigma}_Y \widehat{G}_o| - n \log |\widehat{\Sigma}_{\text{res}}| \\ &= n \log |\widehat{G}^T \widehat{\Sigma}_{\text{res}} \widehat{G}| + n \log |\widehat{G}_o^T \widehat{\Sigma}_{\text{res}} \widehat{G}_o| - n \log |\widehat{\Sigma}_{\text{res}}| \\ &\quad + n \log |I_p + \widehat{\Sigma}_X^{1/2} \hat{\beta}_r^T \widehat{G}_o (\widehat{G}_o^T \widehat{\Sigma}_{\text{res}} \widehat{G}_o)^{-1} \widehat{G}_o^T \hat{\beta}_r \widehat{\Sigma}_X^{1/2}| \end{aligned} \quad (12)$$

where $\widehat{\Sigma}_Y = n^{-1} \mathbb{Y}^T \mathbb{Y}$. The second equation in (12) follows by applying the usual expansion of the determinant of a sum of the form $A + BB^T$. To see this,

$$\begin{aligned} |\widehat{G}_o^T \widehat{\Sigma}_Y \widehat{G}_o| &= |\widehat{G}_o^T \widehat{\Sigma}_{\text{res}} \widehat{G}_o + \widehat{G}_o^T \mathbb{Y}^T \mathbb{X} (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y} \widehat{G}_o| \\ &= |\widehat{G}_o^T \widehat{\Sigma}_{\text{res}} \widehat{G}_o + \widehat{G}_o^T \hat{\beta}_r \widehat{\Sigma}_X \hat{\beta}_r^T \widehat{G}_o| \\ &= |\widehat{G}_o^T \widehat{\Sigma}_{\text{res}} \widehat{G}_o| \times |I_p + \widehat{\Sigma}_X^{1/2} \hat{\beta}_r^T \widehat{G}_o (\widehat{G}_o^T \widehat{\Sigma}_{\text{res}} \widehat{G}_o)^{-1} \widehat{G}_o^T \hat{\beta}_r \widehat{\Sigma}_X^{1/2}|. \end{aligned}$$

We bound λ_j from below by further minimizing the first three addends in (12) over $(\widehat{G}, \widehat{G}_o)$. These are minimized globally when the columns of \widehat{G} span any reducing subspace of $\widehat{\Sigma}_{\text{res}}$ and is 0 at the minimum. Thus

$$\begin{aligned} \lambda_j &\geq n \log |I_p + \widehat{\Sigma}_X^{1/2} \hat{\beta}_r^T \widehat{G}_o (\widehat{G}_o^T \widehat{\Sigma}_{\text{res}} \widehat{G}_o)^{-1} \widehat{G}_o^T \hat{\beta}_r \widehat{\Sigma}_X^{1/2}| \\ &= n \log |I_p + \widehat{\Sigma}_X^{1/2} \hat{\beta}_r^T \widehat{\Sigma}_{\text{res}}^{-1/2} \left\{ \widehat{\Sigma}_{\text{res}}^{1/2} \widehat{G}_o (\widehat{G}_o^T \widehat{\Sigma}_{\text{res}} \widehat{G}_o)^{-1} \widehat{G}_o^T \widehat{\Sigma}_{\text{res}}^{1/2} \right\} \widehat{\Sigma}_{\text{res}}^{-1/2} \hat{\beta}_r \widehat{\Sigma}_X^{1/2}| \\ &= n \log(\widehat{A}_{j,n}), \end{aligned} \quad (13)$$

where $\widehat{A}_{j,n}$ is defined implicitly. The quantity $\widehat{\Sigma}_{\text{res}}^{1/2} \widehat{G}_o (\widehat{G}_o^T \widehat{\Sigma}_{\text{res}} \widehat{G}_o)^{-1} \widehat{G}_o^T \widehat{\Sigma}_{\text{res}}^{1/2}$ in (13) is the projection into the column space of $\widehat{\Sigma}_{\text{res}}^{1/2} \widehat{G}_o$. The quantity $\widehat{G}_o^T \hat{\beta}_r \neq 0$ almost surely since $j = 1, \dots, u-1$. As a result, the column space of $\widehat{\Sigma}_{\text{res}}^{-1/2} \hat{\beta}_r \widehat{\Sigma}_X^{1/2}$ in (13) has a nontrivial intersection with the column space of $\widehat{\Sigma}_{\text{res}}^{1/2} \widehat{G}_o$ almost surely. Therefore $\widehat{A}_{j,n} > 1$ almost surely. We can write $n \log(\widehat{A}_{j,n}) = n |O_p(1)|$ and we have the bound

$$e^{-\lambda_j} = e^{-2\{l(\hat{\beta}_j) - l(\hat{\beta}_r)\}} \leq e^{-n \log(\widehat{A}_{j,n})} = e^{-n|O_p(1)|}.$$

Therefore,

$$\begin{aligned} \log(w_j) &\leq b_u - b_j \\ &= -2\{l(\hat{\beta}_u) - l(\hat{\beta}_r)\} + 2\{l(\hat{\beta}_j) - l(\hat{\beta}_r)\} + \{k(u) - k(j)\} \log(n) \\ &= |O_p(1)| - \lambda_j + \{k(u) - k(j)\} \log(n) \\ &\leq |O_p(1)| - n |O_p(1)| + \{k(u) - k(j)\} \log(n) = -n |O_p(1)| \end{aligned} \quad (14)$$

and we see that $\sqrt{n}w_j \leq \sqrt{n}e^{-n|O_p(1)|}$ for $j = 1, \dots, u-1$.

Define \widehat{G}_o^* to be the estimate of G_o obtained from the starred data and let

$$\begin{aligned} A_{j,n}^* &= |I_p + \widehat{\Sigma}_X^{1/2} \widehat{\beta}_r^{*T} \widehat{G}_o^* \left(\widehat{G}_o^{*T} \widehat{\Sigma}_{\text{res}}^* \widehat{G}_o^* \right)^{-1} \widehat{G}_o^{*T} \widehat{\beta}_r^* \widehat{\Sigma}_X^{1/2}| \\ &= |I_p + \widehat{\Sigma}_X^{1/2} \widehat{\beta}_r^{*T} \widehat{\Sigma}^{*-1/2} \left\{ \widehat{\Sigma}^{*1/2} \widehat{G}_o^* \left(\widehat{G}_o^{*T} \widehat{\Sigma}_{\text{res}}^* \widehat{G}_o^* \right)^{-1} \widehat{G}_o^{*T} \widehat{\Sigma}^{*1/2} \right\} \widehat{\Sigma}^{*-1/2} \widehat{\beta}_r^* \widehat{\Sigma}_X^{1/2}| \end{aligned} \quad (15)$$

The same logic that applied to $\widehat{A}_{j,n}$ applies to $A_{j,n}^*$. The quantity $\widehat{\Sigma}^{*1/2} \widehat{G}_o^* \left(\widehat{G}_o^{*T} \widehat{\Sigma}_{\text{res}}^* \widehat{G}_o^* \right)^{-1} \widehat{G}_o^{*T} \widehat{\Sigma}^{*1/2}$ in (15) is the projection onto the column space of $\widehat{\Sigma}^{*1/2} \widehat{G}_o^*$. The quantity $\widehat{G}_o^{*T} \widehat{\beta}_r^* \neq 0$ almost surely since $j = 1, \dots, u-1$. As a result, the column space of $\widehat{\Sigma}^{*-1/2} \widehat{\beta}_r^* \widehat{\Sigma}_X^{1/2}$ in (15) has a nontrivial intersection with the column space of $\widehat{\Sigma}^{*1/2} \widehat{G}_o^*$ almost surely. Therefore $A_{j,n}^* > 1$ almost surely. The steps in (14), applied to the starred data, yields

$$\sqrt{n}w_j^* \leq \sqrt{n}e^{-n|O_p(1)|}. \quad (16)$$

Thus,

$$\begin{aligned} \sqrt{n} \|w_j^* \text{vec}(\widehat{\beta}_j^*) - w_j \text{vec}(\widehat{\beta}_j)\| &\leq \sqrt{n} \|w_j^* \text{vec}(\widehat{\beta}_j^*)\| + \sqrt{n} \|w_j \text{vec}(\widehat{\beta}_j)\| \\ &\leq \sqrt{n} e^{-n|O_p(1)|} \|\text{vec}(\widehat{\beta}_j^*)\| + \sqrt{n} e^{-n|O_p(1)|} \|\text{vec}(\widehat{\beta}_j)\| \\ &= 2O_p(1) \sqrt{n} e^{-n|O_p(1)|} \end{aligned}$$

for $j = 1, \dots, u-1$ where $\|\text{vec}(\widehat{\beta}_j)\|$ and $\|\text{vec}(\widehat{\beta}_j^*)\|$ are both $O_p(1)$ just as in the $j = u+1, \dots, r$ case. Combining all of these term yields the $2(u-1)O_p(1)\sqrt{n}e^{-n|O_p(1)|}$ order in (8) in Eck and Cook (2017). This completes the proof when $j = 1, \dots, u-1$.

The final case is when $j = u$. Let $E_n = \sum_{i \neq u}^r e^{b_u - b_i}$. We can write $w_u = \frac{1}{1+E_n} = 1 - \frac{E_n}{1+E_n}$. The term $E_n = O_p(n^{-p})$ since $e^{-n|O_p(1)|} = O_p(n^{-p})$. Therefore

$$\begin{aligned} \sqrt{n}w_u^* \text{vec}(\widehat{\beta}_u^*) &= \sqrt{n} \left(1 - \frac{E_n}{1+E_n} \right) \text{vec}(\widehat{\beta}_u^*) \\ &= \sqrt{n} \text{vec}(\widehat{\beta}_u^*) + O_p \left\{ n^{(1/2-p)} \right\}, \\ \sqrt{n}w_u \text{vec}(\widehat{\beta}_u) &= \sqrt{n} \left(1 - \frac{E_n}{1+E_n} \right) \text{vec}(\widehat{\beta}_u) \\ &= \sqrt{n} \text{vec}(\widehat{\beta}_u) + O_p \left\{ n^{(1/2-p)} \right\}. \end{aligned}$$

Adding the previous results over j to form $\sqrt{n} \{ \text{vec}(\widehat{\beta}_w^*) - \text{vec}(\widehat{\beta}_w) \}$ yields the result given in (8) in Eck and Cook (2017). This completes the proof. \square

References

Eck, D. J. and Cook, R. D. (2017). Weighted envelope estimation to handle variability in model selection. *Submitted*.